UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

19|20

Trabajo fin de grado

Information search using random walks on graphs.

Emilio Samuel Aced y Fuentes

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

excelencia Campus Internacional UAM CSIC

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Teoría de grafos

TRABAJO FIN DE GRADO

Information search using random walks on graphs.

Autor: Emilio Samuel Aced y Fuentes
Tutor: Simone Santini

julio 2020

Emilio Samuel Aced y Fuentes
Information search using random walks on graphs.

Emilio Samuel Aced y Fuentes
C\ Francisco Tomás y Valiente Nº 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mi tutor, Adriana y mis padres

No fuisteis creados para vivir como bestias sino para seguir la virtud y la sabiduría.
Fatti non foste a viver come bruti ma per seguir virtute e canoscenza.

Dante Alighieri

Agradecimientos

A mi tutor, Simone Santini, sin el cual este trabajo no se hubiera podido llevar a cabo, por su paciencia en mis momentos de dudas e incertidumbre, guía cuando perdía el rumbo y dedicación en toda circunstancia.

A mi madre por todo lo que hacen las madres y no siempre tenemos en cuenta y brindarme sabio consejo aunque no la escuchara.

A Adriana por ser un faro de tranquilidad y paz en las mareas más turbulentas y levantarme en mis momentos más bajos.

A Judith por todas esas horas de estudio que hemos compartido y que por fin han dado sus frutos.

Por último y no menos importante, a mi padre por instruirme en la vida e inculcarme responsabilidad y sentido común, siempre presente en mis pensamientos.

Resumen

Este trabajo recoge el análisis de diversos tipos de difusión. Repasamos los conceptos más relevantes de los grafos y paseos aleatorios para después estudiar en profundidad el algoritmo Page Rank y los caminos de Lévy. La contribución más destacada de este proyecto es el estudio de la teoría espectral de la matriz de transición del algoritmo de Larry Page. Una vez hecho esto, procedemos a estudiar ambos de forma empírica obteniendo algunas de las magnitudes más relevantes de ambos y finalizamos con un análisis de como de bien modelizan el comportamiento de un usuario en redes sociales y, en consecuencia, la difusión de las publicaciones en estos centros de intercambio de opiniones y noticias.

Palabras clave

Teoría de grafos, paseos aleatorios, teoría espectral, difusión, Page Rank, paseos de Lévy

Abstract

In this project we analise two types of diffusion, we review the most relevant concepts of graphs and random walks. After that, there is a detailed examination of the Page Rank algorithm and Lévy flights. The highlight of this work is an exahustive chapter in which we deveop a spectral theory of the transistion matrix of the Page Rank algorithm. Then, we empirically study and validate the most relevant results of our theoretical study. InFinally, we discuss which is the most suitable to model the behaivour of a social network user and, as a consecuence of this, the diffusion of news and opinions in this sites.

Keywords

Graph theory, diffusión, random walks, spectral theory, Page Rank, Lévy flights

ÍNDICE

Listas

Lista de ecuaciones

Lista de figuras

1

Introducción

Actualmente nos encontramos en un mundo tremendamente globalizado, en el cual, cuando se produce cualquier tipo de interacción persona a persona es muy posible que cada uno se lleve algo del otro. Esto que nos llevamos, pueden ser noticias, leyendas urbanas, opiniones o incluso el resfriado del niño que tienes al lado en el autobús. Este fenómeno de trasmisión se denomina difusión y tiene suma relevancia en la era en la que vivimos, un ejemplo claro es la actual crisis que estamos sufriendo a raíz del virus Covid-19, los epidemiólogos estudian su difusión con el fin de ver qué medidas son necesarias para frenarlo.

En este trabajo nos centraremos en la difusión de la información en las llamadas ”redes sociales”. Las redes sociales existen por lo menos desde que existe el lenguaje. El intercambio de información en los mercados, el cotilleo, el correo, el teléfono y, muy prominente hoy en día, internet: todos estos se pueden ver como ejemplos de redes sociales de intercambio de información. Y, desde que existen las redes sociales, existe en ellas la propagación, intencional o no, de noticias y rumores falsos.

Las redes sociales tienen características a veces sorprendentes que, empezando en el Siglo XX, han sido estudiadas ya sea experimentalmente o por medio de modelos matemáticos. Un ejemplo es el experimento de pequeños mundos [1], realizado en los años 60, que consistía en hacer llegar una carta desde Nebraska a Massachusetts bajo una restricción **:** el remitente solo podía poner como destinatario a una alguien que conociera en persona. El resultado fue sorprendente: la mayoría de las cartas llegaron al destinatario tras haber sido enviadas a menos de seis personas diferentes. Esto resultó en el famoso principio de los seis grados de separación [2], que afirma que cualquier persona puede llegar a cualquier otra a través de una cadena de longitud 6 relaciones.

Los seres humanos no hemos tenido en nuestra historia un volumen de datos tan grande como en los últimos veinte años. Esto, unido a que la información que tenemos hoy en día está estructurada, nos ofrece la oportunidad de estudiar su difusión de una forma mucho más eficiente que nuestros predecesores. Internet nos provee cada día de millones de historias, noticias y entretenimiento y todo eso queda almacenado, el ejemplo por excelencia son las redes sociales que se pueden ver cómo un grafo donde las personas son los vértices y las relaciones de amistad o seguimiento son las aristas. En ellas, los usuarios publican un sin fin de información que se propagará a sus seguidores y estos,a

su vez, pueden redistribuirla. Para modelizar la difusión podemos hacerlo de dos maneras, la primera, esperar a que la información se difunda o, la segunda, movernos nosotros, en este trabajo, usaremos paseos aleatorios sobre grafos que cumplan ciertas condiciones que nos aproximen una red social. Nos fijaremos especialmente en dos, los paseos de Lévy y el Page Rank en los que veremos sus diferencias con el objetivo de compararlos en sus distintas propiedades.

En relación a la difusión, analizaremos dos tipos de paseos aleatorios que la reproducen, estos son el algoritmo Page Rank desarrollado por Larry Page [3] en 1996 como parte del proyecto de un nuevo motor de búsqueda y los paseos de Lévy [4] en los cuales la probabilidad de saltar a un nodo depende de la distancia a la que se encuentre respecto a dónde nos encontremos. Nuestro trabajo se desarrollará en torno al Page Rank ya que modeliiza mejor el comportamiento de un usuario de internet que busca información. El usuario normalmente ejecuta una acción de uno de tos tipos posibles: (1) selecciona un enlace en una página que está leyendo, llegando así a otra página o (2) usa un mecanismo alternativo (p.ej. un motor de búsqueda) para llegar a una página que no tiene relación estructural con la de partida. Como veremos, estas dos acciones corresponden a las dos posibilidades de un paseo aleatorio Page Rank: desde un nodo se puede saltar (1) a uno de los vecinos o (2) a un nodo elegido al azar.

2

Base teórica

En este capítulo daremos la teoría y los detalles técnicos a partir de los cuales hemos generado nuestras simulaciones empíricas. Empezaremos con unas definiciones básicas de teoría de grafos, después definiremos formalmente que es un paseo aleatorio y estudiaremos en detalle los dos que analizaremos en profundidad.

2.1. Grafo

Un grafo de N nodos se denota como $\mathcal{G} = (V, E)$ donde $V = \{1, \ldots, N\}$ es el conjunto de vértices que almacenan información y E el de aristas que lo componen, $E \subseteq V \times V$ Existen varios tipos de grafos

- Si las relaciones que componen las aristas son simétricas, lo denominamos grafo no dirigido, si por el contrario estas relaciones no cumplen esa propiedad se considera que el grafo no es dirigido.
- Si podemos separar el conjunto de vértices en dos que no tienen ninguna arista que los conecte disjuntos diremos que es no es conexo, si esto no es posible diremos que es conexo.

Con el fin de facilitar la comprensión de las secciones venideras es necesario dar algunas definiciones que se utilizarán más adelante.

Definiciones sobre nodos:

- Vecindad de un nodo ($\mathcal{V}(u) = \{v \,|\, (u, v) \in E\}$): conjunto de todos los vecinos de un nodo.
- Nodo vecino($v(u)$): un nodo v es vecino de un nodo u si $(u, v) \in E$
- Grado de un nodo ($g(u)$): coincide con el número de aristas que salen de él. Se puede ver como el cardinal de la vecindad del nodo u ($|\mathcal{V}(u)|$)

Por su parte para facilitar el trabajo con los grafos, existen matrices y magnitudes que nos describen el grafo sin necesidad de verlo explícitamente:

- Camino del nodo i al nodo j: $c(i, j) = [u_1, \ldots, u_m]$ tal que $u_h \in V$, $u_1 = i$, $u_m = j$, $(u_h, u_{h+1}) \in E$, $i = 1, \ldots, m - 1$.
  Nota: esta claro que no tiene que existir un único camino entre dos nodos, el conjunto de todos ellos lo denotamos como $C(v_i, v_j)$.

- Matriz de adyacencia: nos da las aristas que conectan los nodos. Tiene un 1 en la entrada (i, j) si existe una arista que nos una el nodo v_i con el nodo v_j. Es decir,

$$A_{ij} = \begin{cases} 1 \text{ si } \exists (v_i, v_j) \in E \\ 0 \text{ en caso contrario} \end{cases} \tag{2.1}$$

- Matriz de distancias: nos da el número de saltos mínimo necesario para ir del nodo v_i al nodo v_j, es decir, el cardinal del camino más corto. De tal forma que

$$S_{i,j} = \min \{|c(v_i, v_j)| \text{ tal que } c(v_i, v_j) \in C(v_i, v_j)\} \tag{2.2}$$

- Matriz de grados: es una matriz diagonal definida como

$$G_{ij} = \begin{cases} g(v_i) \text{ si } i = j \\ 0 \text{ si } i \neq j \end{cases} \tag{2.3}$$

Debido a que nuestro trabajo va principalmente orientado a las redes sociales necesitábamos un grafo que emulara sus propiedades, por esto, utilizamos el modelo definido en [5]. El grafo se crea con el algoritmo siguiente:

1.– Cada vez que llega un nodo u (se crea).

2.– Se le asocia un periodo de vida a que sigue una distribución exponencial.

3.– Al nodo u se le añade una primera arista al nodo v con probabilidad proporcional a su grado.

4.– Un nodo u de grado d con tiempo de espera δ espera $\delta^{-\alpha} e^{-\beta d \delta}$ para su siguiente adición de arista.

5.– Si su tiempo de vida no ha terminado, cuando cumple el tiempo de espera generado crea una arista que cierre un triángulo de nodos.

6.– Si su tiempo de vida ha concluido, ese nodo no volverá a generar ninguna arista.

2.2. Paseos aleatorios estándar

Esta sección basada en [6] usaremos grafos $\mathcal{G} = (V, E)$ no dirigidos ya que como veremos más adelante vamos a utilizar que la matriz de adyacencia A sea simétrica. Un paseo aleatorio sobre un grafo simula una difusión sobre el mismo. Para esto, fijamos un nodo sobre el que iniciaremos nuestro camino en el tiempo $t = 0$ y para a la siguiente unidad de tiempo, nos habremos movido a otro nodo del grafo, este movimiento viene dado por la probabilidad de estar en el nodo j al tiempo t habiendo empezado el movimiento en i $(P_{ij}(t))$. Un camino aleatorio tiene asociada una matriz de transición W con componentes ω_{ij} definidos como:

$$\omega_{ij} = \begin{cases} \frac{1}{d(j)} \text{ si } \exists (v_i, v_j) \in E \\ 0 \text{ en caso contrario} \end{cases} \tag{2.4}$$

tal y como hemos definido la matriz W cada componente ω_{ij} es la probabilidad de moverse desde

el nodo j al nodo i, el nodo inicial del paseo aleatorio se puede elegir de varias formas, pero este fenómeno se puede modelizar a través de un vector de probabilidad $\boldsymbol{p(0)}^T = (p_1(0), p_2(0), \ldots, p_N(0))$ en el que la coordenada i-ésima da la probabilidad de que el nodo i sea el inicial. A partir de estas dos definiciones podemos calcular el vector de probabilidad al tiempo t ($\boldsymbol{p(t)}$) de la siguiente forma

$$\boldsymbol{p(t)} = W\boldsymbol{p(t-1)} = W^t\boldsymbol{p(0)}. \tag{2.5}$$

Una pregunta lógica es plantearse si el camino aleatorio llega alguna vez a un estado estacionario. Este estado se caracteriza por que el vector de probabilidad que describe la posibilidad de estar en cada nodo ya no depende del tiempo, matemáticamente exigimos que

$$\boldsymbol{p(t)} = \boldsymbol{p(t-1)}, \tag{2.6}$$

bajo la construcción dada en la ecuación 2.5 pedimos que se llega a un vector $\boldsymbol{\pi}$ tal que

$$W\boldsymbol{\pi} = \boldsymbol{\pi}. \tag{2.7}$$

con π_i=$\frac{g(i)}{\sum_{j=1}^{N} g(j)}$.El hecho de que en 2.5 nos quede una matiz elevada a un exponente nos lleva a la siguiente sección.

2.2.1. Teoría espectral

La teoría espectral es el estudio de los autovalores y autovectores de las matrices, nos permite entender qué ocurre cuando multiplicamos por una. Esta sección del trabajo se basa en [6].
Recordemos que un autovector $\boldsymbol{v}$ de una matriz M (operador lineal) es un vector no nulo que, cuando se transforma por la matriz genera un múltiplo del mismo, en otras palabras, no cambian su dirección. Un autovalor es el escalar λ por el cual multiplicamos el autovector y obtenemos el vector generado por la multiplicación del autovector por la matriz, esto es,

$$M\boldsymbol{v} = \lambda\boldsymbol{v} \tag{2.8}$$

Teorema 2.1. *Para toda matriz simétrica $M_{N\times N}$ existe una base ortonormal de N autovectores $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_N}$ y N autovalores $\lambda_1, \lambda_2, \ldots, \lambda_N$ tal que*

$$M\boldsymbol{v_i} = \lambda_i\boldsymbol{v_i} \tag{2.9}$$

para todo $i = 1, 2, \ldots, N$.

La ortonormalidad que nos asegura el teorema 2.1 nos da una forma de expresar cualquier vector $\boldsymbol{u}$ en esta base a través de la combinación lineal

$$\boldsymbol{u} = \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) \boldsymbol{v_i} \tag{2.10}$$

donde el término $(\boldsymbol{v_i}^T \boldsymbol{u})$ es un escalar y, por ello el coeficiente del autovector $\boldsymbol{v_i}$ de la expresión de $\boldsymbol{u}$ en la base $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_N}\}$. Obtenemos así una nueva fórmula de multiplicación de matrices,

$$M\boldsymbol{u} = M \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) \boldsymbol{v_i} = \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) M\boldsymbol{v_i} = \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) \lambda_i \boldsymbol{v_i}. \tag{2.11}$$

La ventaja de la nueva expresión es que calcular la potencia de una matriz por un vector es mucho más sencillo,

$$M^k\boldsymbol{u} = M \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) \boldsymbol{v_i} = M \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) \boldsymbol{v_i} = \sum_{i=1}^{N} (\boldsymbol{v_i}^T \boldsymbol{u}) \lambda_i^k \boldsymbol{v_i}. \tag{2.12}$$

Todos estos resultados los aplicaremos a nuestra matriz de transición W que aunque no sea diagonal podemos transformarla de forma que sí lo sea. Las matrices que tienen esta propiedad, se las denomina matrices semejantes a una matriz simétrica. Veamos la transformación de W, si tomamos la matriz G cuyo elemento i-ésimo de la diagonal es el grado del nodo i, la matriz $G^{\frac{1}{2}}$ será la matriz diagonal con elementos $\sqrt{g(i)}$ en la diagonal, entonces

$$G^{-\frac{1}{2}} W G^{\frac{1}{2}} = G^{-\frac{1}{2}} (AG^{-1}) G^{\frac{1}{2}} = G^{-\frac{1}{2}} A G^{-\frac{1}{2}} \tag{2.13}$$

que es simétrica y la llamaremos a partir de aquí matriz M, esta nueva matriz tiene los mismos autovalores que W ya que por definición de autovector

$$\lambda_i \boldsymbol{v_i} = M\boldsymbol{v_i} = G^{-\frac{1}{2}} W G^{\frac{1}{2}} \boldsymbol{v_i} \implies \lambda_i (G^{\frac{1}{2}} \boldsymbol{v_i}) = W (G^{\frac{1}{2}} \boldsymbol{v_i}). \tag{2.14}$$

Gracias a esto, ahora estamos en condiciones de aplicar la teoría espectral a la matriz W de tal forma que

$$W^k\boldsymbol{u} = (G^{\frac{1}{2}} M G^{-\frac{1}{2}})^k \boldsymbol{u} = G^{\frac{1}{2}} M^k G^{-\frac{1}{2}} \boldsymbol{u} = \sum_{i=1}^{N} \lambda_i^k \boldsymbol{v_i} (\boldsymbol{v_i}^T G^{-\frac{1}{2}} \boldsymbol{u}), \tag{2.15}$$

ahora para ver cual será la distribución estacionaria utilizamos el siguiente teorema [6].

Teorema 2.2. *Sea W la matriz de transición de un grafo conexo, entonces todos los autovalores λ_i estarán entre 1 y -1 y el autovalor 1 tendrá multiplicidad 1.*

Demostración. Esto se deriva de que $\boldsymbol{\pi}$ es un autovector con autovalor $\lambda = 1$ y $\pi_i \geq 0$ y el teorema de

Perron-Frobenius. □

Nota: A partir de aquí y sin pérdida de generalidad nombraremos a los autovalores $1 = \lambda_1 < \lambda_2 < \ldots < \lambda_N \leq -1$ y a sus autovectores asociados $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_N}$.

Tras este resultado y, asumiendo que no existe el autovalor -1, es fácil ver que la distribución estacionaria será el sumando asociado al autovalor 1 ya que en el último término de la expresión 2.15 todos los autovalores salvo el 1 tenderán a 0 a un t muy grande y al hacer el límite cuando $t \to \infty$ dará

$$\lim_{t\to\infty} W^t \boldsymbol{u} = \boldsymbol{v_i}(\boldsymbol{v_i}^T G^{-\frac{1}{2}} \boldsymbol{u}) \tag{2.16}$$

En la sección 2.5 veremos en qué casos podemos encontrar un autovalor -1 en la matriz W y un ti‘p de àseo aleatorio que si que converge a una distribución estacionaria.

2.3. Paseos de Lévy

Los resultados de esta sección se basan en [6]. Consideremos un grafo no dirigido $\mathcal{G} = (V, E)$, la matriz de adyacencia A, la matriz G de grados y la matriz S definidas en la sección 2.2. Teniendo en cuenta la ecuación discreta del tiempo,

$$p_{ij}(t) = \sum_{m=1}^{N} p_{im}(t)\omega_{ij} \tag{2.17}$$

donde ω_{ij} es la probabilidad de transición de i a j definida como

$$\omega_{ij} = \frac{S_{ij}^{-\alpha}}{\sum_{l\neq i}^{N} S_{il}^{-\alpha}}, \alpha \in [0, \infty) \tag{2.18}$$

y $p_{ij}(t)$ es la probabilidad de estar en en nodo j al tiempo t habiendo empezado el camino en el nodo i, esta modelización de un caminante aleatorio no solo te permite visitar los nodos vecinos como ocurría en el paseo aleatorio estándar sino que permite visitar cualquier nodo del grafo con una probabilidad que depende del parámetro α y decae cuanto más lejos esté el nodo destino. Este parámetro α afecta la forma en la que se recorre el grafo, es relevante destacar los dos casos límite.

Si tomamos el límite cuando $\alpha \to \infty$

$$\omega_{ij} = \lim_{\alpha\to\infty} \frac{S_{il}^{-\alpha}}{\sum_{l\neq i}^{N} S_{il}^{-\alpha}} = \frac{A_{ij}}{G_{ii}} \tag{2.19}$$

y tenemos así la traspuesta de la matriz de transición de un paseo aleatorio estándar, si $\alpha = 0$ tenemos

$$\omega_{ij} = \frac{1}{\sum_{l\neq i}^{N} 1} = \frac{1-\delta_{ij}}{N-1} \tag{2.20}$$

con δ_{ij} la delta de Kronecker, es relevante ver que la probabilidad de saltar a cualquier nodo es la misma, esto se puede ver como que el paseo de Lévy con $\alpha = 0$ dará la sensación de estar recorriendo un grafo completo. En el resto de casos, $0 < \alpha < \infty$ estamos haciendo posible que el caminante salte a un nodo a distancia mayor que 1 con una probabilidad proporcional a la distancia y que decae como una potencia.

Consideremos el siguiente problema, si empezamos en el tiempo $t = 0$ con la distribución estacionaria, por iteraciones en la fórmula 2.17 la probabilidad p_{ij} toma la forma

$$p_{ij}(t) = \sum_{j_1,\ldots,j_{t-1}} \omega_{ij_1}\omega_{j_1j_2}\cdots\omega_{j_{t-1}j}. \tag{2.21}$$

Si definimos ahora la cantidad $\psi_i^{(\alpha)} = \sum_{l\neq i} S_{il}^{-\alpha}$ nos da que

$$\omega_{ij} = \frac{\psi_j^{(\alpha)}}{\psi_i^{(\alpha)}}\omega_{ji} \tag{2.22}$$

que junto a la ecuación 2.21 nos da lo condición de equilibrio

$$\psi_i^{(\alpha)} p_{ij}(t) = \psi_j^{(\alpha)} p_{ji}(t). \tag{2.23}$$

Para obtener la distribución estacionaria π_i, tenemos que tomar el límite de la probabilidad de ocupación del nodo i cuando el tiempo tiende a infinito e interpretando π_i como la media temporal de $\lim_{t\to\infty} p_{ij}(t)$ nos queda junto con la condición 2.23

$$\pi_i = \frac{\psi_i^{(\alpha)}}{\sum_{l=1}^{N} \psi_l^{(\alpha)}} \tag{2.24}$$

esta probabilidad estacionaria depende a su vez del parámetro α, cuando $\alpha \to \infty$, $\quad \pi_i = \frac{g(i)}{\sum_{j=1}^{N} g(j)}$ que queda igual que en un paseo aleatorio estándar y si $\alpha = 0$, $\quad \pi_i = \frac{1}{N}$.
Desde aquí, vamos a estudiar el tiempo medio para llegar por primera vez al nodo j desde el nodo i, expresamos la probabilidad de ocupación como

$$p_{ij}(t) = \delta_0\delta_{ij} + \sum_{t_0=0}^{T} p_{jj}(t-t_0)F_{ij}(t_0) \tag{2.25}$$

donde $F_{ij}(t)$ es la probabilidad de haber encontrado el nodo j al tiempo t, si a esta fórmula le aplicamos la transformada de Laplace tenemos una igualdad que nos permite calcular $\tilde{F}_{ij}(s)$

$$\tilde{F}_{ij}(s) = \frac{\tilde{p}_{ij}(s) - \delta_{ij}}{\tilde{p}_{jj}(s)} \tag{2.26}$$

ahora, para redes finitas el tiempo medio para llegar por primera vez al nodo j desde el nodo i se obtiene a partir de la igualdad

$$\langle T_{ij} \rangle = \sum_{t=0}^{\infty} t F_{ij}(s) = -\frac{\mathrm{d}\tilde{F}_{ij}(0)}{\mathrm{d}s} \tag{2.27}$$

si nombramos al n-ésimo momento como $R_{ij}^{(n)} = \sum_{t=0}^{\infty} t^n (p_{ij} - \pi_j)$ la expansión de $\tilde{p}_{ij}$ es

$$\tilde{p}_{ij}(s) = \frac{\pi_j}{1 - e^{-s}} + \sum_{n=0}^{\infty} (-1)^n R_{ij}^{(n)} \frac{s^n}{n!} \tag{2.28}$$

que al introducirlo en 2.26, cambiar el signo, derivar respecto a s y evaluar en 0 nos da $\langle T_{ij} \rangle$

$$\langle T_{ij} \rangle = \frac{R_{ij}^{(0)} - R_{ij}^{(0)} + \delta_{ij}}{\pi_j} \tag{2.29}$$

Para terminar, hallamos el tiempo medio entre nodos una vez alcanzada la distribución estacionaria $\langle T \rangle$ como la media de los $\langle T_{ij} \rangle$

$$\langle T \rangle = \sum_{j \neq i}^{N} \langle T_{ij} \rangle \pi_j = \sum_{m=1}^{N} R_{mm}^{(0)} \tag{2.30}$$

Nos falta, por tanto, calcular los momentos, esto lo veremos en la siguiente sección.

2.3.1. Teoría espectral de los paseos de Lévy

Para poder calcular varias cantidades definidas en la sección 2.3 necesitamos hallar los momentos $R_{ij}^{(0)}$ y, por la definición de estos, tenemos que hallar $p_{ij}(t)$, esto lo conseguiremos a través de la teoría espectral. Sabemos que en forma matricial, la ecuación de un paseo de Lévy tiene la siguiente forma,

$$\boldsymbol{p(t)}^T = \boldsymbol{p(0)}^T W^t \tag{2.31}$$

Por la definición de la matriz W podemos afirmar que,

$$p_{ij}(t) = \boldsymbol{e_i}^T W^t \boldsymbol{e_j} \tag{2.32}$$

donde el $\boldsymbol{e_i}$ es el i-ésimo vector de la base canónica. Gracias a los resultados de [4] sabemos que los autovalores $\lambda_1, \lambda_2, \ldots, \lambda_N$ de la matriz de transición W cumplen $|\lambda_i| \in (0,1)\ \forall i = 2, \ldots, N$, además, existe un autovalor denotado por $\lambda_1 = 1$, estos, tienen autovectores asociados $\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_N}$ de tal forma que

$$W\boldsymbol{z_i} = \lambda_i \boldsymbol{z_i} \tag{2.33}$$

Estos autovectores forman la matriz Z con elementos $Z_{ij} = \boldsymbol{e_i}^T \boldsymbol{z_j}$ con $\boldsymbol{z_j}$ el j-ésimo autovector derecho, esta matriz es invertible y su inversa se denota por Z^{-1} con elementos $Z^{-1}_{ij} = \boldsymbol{\zeta_i}^T \boldsymbol{e_j}$con $\boldsymbol{\zeta_i}$ el i-ésimo autovector izquierdo, gracias a esto tenemos una nueva forma de representar $p_{ij}(t)$

$$p_{ij}(t) = \boldsymbol{e_i}^T (Z\Lambda^t Z^{-1}) \boldsymbol{e_j} = \sum_{l=1}^{N} \lambda_l^t (\boldsymbol{e_i}^T \boldsymbol{z_l})(\boldsymbol{\zeta_l}^T \boldsymbol{e_j}) \tag{2.34}$$

desde esta expresión y gracias al rango donde están definidos los autovalores tenemos la distribución estacionaria,

$$\pi_i = (\boldsymbol{e_i}^T \boldsymbol{z_1})(\boldsymbol{\zeta_1}^T \boldsymbol{e_i}) \tag{2.35}$$

aplicando 2.34 en 2.29 obtenemos

$$R_{ij}^{(0)} = \sum_{l=2}^{N} \frac{1}{1-\lambda_l} (\boldsymbol{e_i}^T \boldsymbol{z_l})(\boldsymbol{\zeta_l}^T \boldsymbol{e_i}) \tag{2.36}$$

y nos permite hallar

$$\langle T_{ij} \rangle = \sum_{l=2}^{N} \frac{1}{1-\lambda_l} \frac{(\boldsymbol{e_i}^T \boldsymbol{z_l}) - (\boldsymbol{\zeta_l}^T \boldsymbol{e_i})}{(\boldsymbol{e_i}^T \boldsymbol{z_1})(\boldsymbol{\zeta_1}^T \boldsymbol{e_i})} \tag{2.37}$$

y el tiempo medio entre nodos

$$\langle T \rangle = \sum_{l=2}^{N} \frac{1}{1-\lambda_l} \tag{2.38}$$

2.4. Algoritmo de PageRank

El algoritmo de Page Rank de Google es utilizado a diario por millones de personas, sin embargo, no sabemos a ciencia cierta su comportamiento intrínseco, por ello, en el capítulo 3 veremos el rendimiento empírico de dos de sus magnitudes relevantes y cómo calcularlas de forma teórica, tiempo medio entre nodos y nodos visitados en función del tiempo. A modo introductorio veamos la matriz de transición, esta matriz $\overline{W}$ depende de un parámetro $\beta \in [0, 1]$ que da probabilidad de salto a nodos vecinos. La construcción de $\overline{W}$ es la siguiente,

$$\overline{W} = \frac{1-\beta}{N} \mathbf{1} + \beta W \tag{2.39}$$

con W definida en la sección 2.2 y $\mathbf{1}$ la matriz N por N llena de unos. Si observamos la fórmula 2.39 vemos que, en los casos límite hay cierto paralelismo con un paseo de Lévy. Si $\beta = 1$ tenemos

$$\overline{W} = W \tag{2.40}$$

y si $\beta = 0$ queda

$$\overline{W} = \frac{1}{N}. \tag{2.41}$$

Vemos que en el caso $\beta = 1$ coincide con la matriz de transición del camino de Lévy con $\alpha = \infty$ y responde al comportamiento de un paseo aleatorio estándar. En el caso $\beta = 0$ es muy parecida al caso $\alpha = 0$ de un camino de Lévy respecto a que el caminante aleatorio puede ir a cualquier nodo del grafo con la misma probabilidad sin importar la distancia a la que se encuentre, sin embargo, en este caso de $\overline{W}$ da la posibilidad de que el caminante aleatorio no se mueva de dónde está al tiempo $t+1$, aunque es ínfima y esperamos que no afecte de forma significativa. El estudio de las característica de los caminos aleatorios de tipo Page Rank es uno de los objetivos del presente trabajo, y lo consideraremos en el capítulo siguiente.

2.5. Grafos bipartitos

Un grafo bipartito $\mathcal{G} = (V, E)$ es aquel en el cual podemos escoger dos conjuntos $U \subset V$ y $W \subset V$ de vértices que cumplan la condición $\forall u_1, u_2 \in U, \forall w_1, w_2 \in W \quad \nexists e \in E \mid e = (u_1, u_2)$ ni $e = (w_1, w_2)$ y que además,

- $U \cap W = \emptyset$.
- $U \cup W = V$.

En la figura 2.1 vemos un ejemplo de un grafo que cumple todas las condiciones mencionadas. Estas restricciones hacen que un grafo bipartito de $n + m$ nodos se puedan extraer dos subconjuntos de n y m nodos y, etiquetando los nodos del primer subconjunto como $1, \ldots, n$ y los del segundo como $n + 1, \ldots, n + m$, se tiene que la matriz de adyacencia es de la forma

$$A = \begin{pmatrix} \mathbf{0_{m\times n}} & B \\ B^T & \mathbf{0_{n\times m}} \end{pmatrix}$$

donde $\mathbf{0_{m\times n}}$ es la matriz de ceros de dimensión $m \times n$. Estos grafos tienen una particularidad, a la hora de realizar paseos aleatorios sobre ellos no existe convergen a una distribución estacionaria. Esto se prueba en la siguiente sección 2.5.1.

2.5.1. Teoría espectral de los paseos aleatorios en grafos bipartitos

Como ya adelantamos en la sección 2.2.1, una magnitud interesante de los paseos aleatorios es ver la probabilidad de estar en un nodo para un tiempo muy grande. Sin embargo, para un grafo bipartito, esta distribución estacionaria no converge, el motivo por el cual esto ocurre se deriva del siguiente teorema.

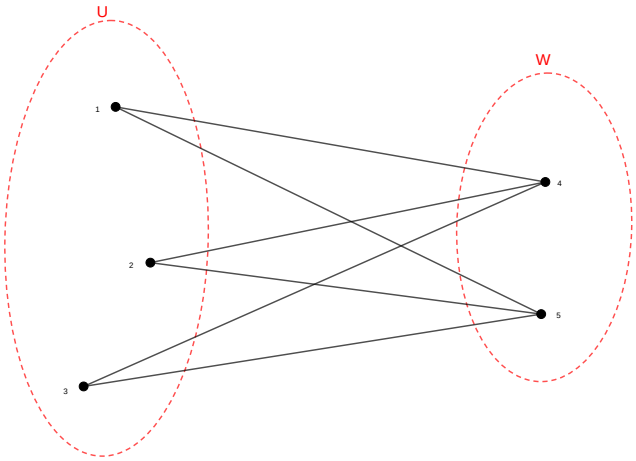

Figura 2.1: Ejemplo de grafo bipartito

Teorema 2.3. *Un grafo es bipartito* $\Leftrightarrow$ *La matriz de transición* W *definida en la sección 2.2 tiene un autovalor* $\lambda = -1$

Demostración. Para la demostración utilizaremos la matriz W^T ya que comparte autovalores con la matriz W.

$\Rightarrow$) Suponiendo que un grafo $\mathcal{G} = (V, E)$ es bipartito, entonces tiene como matriz de adyacencia

$$A = \begin{pmatrix} \mathbf{0} & B \\ B^T & \mathbf{0} \end{pmatrix} \implies W^T = T = \begin{pmatrix} \mathbf{0} & W_1 \\ W_2 & \mathbf{0} \end{pmatrix}$$

con W_1 de dimensión $n \times m$ y W_2 de dimensión $m \times n$, como por construcción de W sabemos que

$$\sum_{j=1}^{|V|} T_{i,j} = \sum_{j=1}^{|g(i)|} \frac{1}{|g(i)|} = \frac{|g(i)|}{|g(i)|} = 1 \quad \forall j = 1, 2, \ldots, N \tag{2.42}$$

podemos afirmar

$$T\mathbf{1_{n+m}} = \mathbf{1_{n+m}} \tag{2.43}$$

con $\mathbf{1_{n+m}}$ el vector de $n + m$ unos. Por la fórmula de W de grafos bipartitos podemos afirmar

$$W_1\mathbf{1_m} = \mathbf{1_n} \quad W_2\mathbf{1_n} = \mathbf{1_m} \tag{2.44}$$

y tomando el vector

$$\boldsymbol{b} = \begin{pmatrix} \mathbf{1_n} \\ -\mathbf{1_m} \end{pmatrix}$$

junto con las propiedades de T tenemos que al multiplicarlos

$$T\boldsymbol{b} = \begin{pmatrix} \mathbf{0} & W_1 \\ W_2 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{1_n} \\ \mathbf{-1_m} \end{pmatrix} = \begin{pmatrix} -W_1\mathbf{1_m} \\ W_2\mathbf{1_n} \end{pmatrix} = -\boldsymbol{b}$$

por tanto, podemos afirmar que $\boldsymbol{b}$ es un autovector con autovalor asociado $\lambda = -1$.
$\Leftarrow$)Supongamos ahora que W tiene un autovalor -1. Con $\boldsymbol{b}$ su autovector derecho asociado, por tanto,

$$\boldsymbol{c} = T\boldsymbol{b} \tag{2.45}$$

de donde se deriva que $c_j = b_j$. A partir de aquí demostraremos por contradicción que todos los componentes de $\boldsymbol{b}$ tienen el mismo valor absoluto. Supongamos que

$$b_h = \min_j \{b_j\} < \max_j \{b_j\} = b_l \tag{2.46}$$

lo que implica,

$$|c_j| = \left| \sum_{j=1}^{n+m} T_{ij} b_j \right| \leq \sum_{j=1}^{n+m} T_{ij} \, |b_j| < |b_l| \sum_{j=1}^{n+m} T_{i,j} = b_l \Rightarrow\Leftarrow \tag{2.47}$$

Como hemos visto que todos los elementos de $\boldsymbol{b}$ tienen el mismo valor absoluto escogemos el vector cuyas componentes son cantidades unitarias, esto es, $\boldsymbol{b} \in \{1, -1\}_{n+m}$ y, como podemos desplazar las filas y columnas de T, lo hacemos de tal manera que los unos estén en las n primeras posiciones, quedándonos el mismo vector $\boldsymbol{b}$ que en el apartado anterior de la prueba. Tras este ajuste, los primeros n elementos deberán ser

$$c_j = \sum_{j=1}^{k} T_{ij} b_j = \sum_{j=1}^{n} T_{ij} - \sum_{j=n+1}^{n+m} T_{ij} = -b_j = -1 \tag{2.48}$$

Como $T_{ij} \geq 0$ y $\sum_{j=1}^{|V|} T_{ij} = 1$, la única forma de obtener -1 es que $\omega_{i,j} = 0 \quad \forall i = 1, \ldots, n$, lo que nos dice

$$\sum_{i=n+1}^{n+m} T_{ij} = \sum_{i=1}^{n+m} T_{ij} = 1 \tag{2.49}$$

veamos que ocurre si $j = n+1, \ldots, n+m$

$$c_j = \sum_{j=1}^{n+m} T_{ij} b_i = \sum_{j=1}^{n} T_{i,j} - \sum_{j=n+1}^{n+m} T_{ij} = -b_j = 1 \tag{2.50}$$

y gracias a las mismas condiciones,

$$\sum_{j=1}^{n} T_{ij} = \sum_{j=1}^{n+m} T_{ij} = 1 \tag{2.51}$$

para terminar, juntamos las dos condiciones 2.51 y 2.49 quedándonos

$$T_{ij} = 0 \text{ si } (j \leq n \text{ y } i \leq m) \text{ o } (j > n \text{ y } i > m) \tag{2.52}$$

que en forma matricial nos da la fórmula de la matriz de T que concuerda con la de un grafo bipartito.

□

Gracias a este teorema vemos que la expresión de la distribución estacionaria no converge ya que queda si t es par

$$\lim_{t\to\infty} W^t \boldsymbol{p(0)} = \boldsymbol{v_1}(\boldsymbol{v_1}^T G^{-\frac{1}{2}} \boldsymbol{p(0)}) + \boldsymbol{v_N}(\boldsymbol{v_N}^T G^{-\frac{1}{2}} \boldsymbol{p(0)}) \tag{2.53}$$

y si t es impar

$$\lim_{t\to\infty} W^t \boldsymbol{p(0)} = \boldsymbol{v_1}(\boldsymbol{v_1}^T G^{-\frac{1}{2}} \boldsymbol{p(0)}) - \boldsymbol{v_N}(\boldsymbol{v_N}^T G^{-\frac{1}{2}} \boldsymbol{p(0)}) \tag{2.54}$$

que cuando $t \to \infty$ va pegando saltos de una a otra y no convergerá a ningún valor.

Nota: Existen los paseos aleatorios denominados perezosos con matriz de transición

$$\mathcal{W} = \frac{I}{2} + \frac{W}{2} \tag{2.55}$$

que siempre convergen a una distribución estacionaria, además, es relevante destacar que la mitad de las veces no se mueven al incrementar el tiempo una unidad.

3

Estudio del algoritmo Page Rank y paseos de Lévy

Como hemos visto en las secciones anteriores, se pueden apreciar similitudes entre ambos algoritmos. Este hecho nos da la idea de que su comportamiento esté ligado y que en el fondo, podamos ver ambos como dos formas de expresar el mismo tipo de difusión. En esta sección intentaremos responder a las siguientes preguntas:

- ¿Cómo evoluciona el número de nodos visitados en el algoritmo Page Rank?
- ¿Existe alguna función que aproxime el comportamiento de número de nodos alcanzados en función del tiempo?
- ¿Cuál es el tiempo medio entre dos nodos cualesquiera en el Page Rank?

Para responder a estas preguntas daremos resultados teóricos del tiempo medio entre nodos y el número de nodos visitados en función del tiempo y proveeremos gráficas obtenidas a partir de simulaciones con el fin de comparar nuestros resultados teóricos con el mundo real y expondremos las conclusiones a las que hemos llegado.

3.1. Teoría espectral del algoritmo Page Rank

Este tipo de paseo aleatorio tuvo sus inicios en 1996 y fue desarrollado por Larry Page y Sergei Brin [3], lo que más nos llamó la atención de este tipo de difusión es que, aunque se utiliza a nivel global, carecemos de una base teórica que nos construya sus magnitudes, por ello, decidimos estudiarlo y expresar algunas de sus propiedades más relevantes.

3.1.1. Ecuaciones básicas

Sea $\mathcal{G} = (V, E)$ un grafo no dirigido con $|V| = N$, matriz de adyacencia A y $g(i) = \sum_{j=1}^{N} A_{ij}$ el grado del nodo i. Con $\theta = 1 - \beta$ las entradas de la matriz de transición $\overline{W}$ del Page Rank serán

$$\overline{\omega}_{ij} = (1-\theta)\frac{A_{ij}}{g(j)} + \theta\frac{1}{N} \tag{3.1}$$

en la cual vemos que dicho algoritmo a cada paso de tiempo lanza una θ-moneda. Si en el lanzamiento sale cara, saltamos a un nodo aleatorio del grafo y si sale cruz, nos movemos a un vecino del nodo en el

que nos encontremos. Sean las matrices D y W definidas en la sección 2.3.1 tenemos que $W = AG^{-1}$ con elementos $\omega_{ij} = \frac{A_{i,j}}{g(j)}$ y, dada una distribución de probabilidad inicial $\boldsymbol{p(0)}$, denotaremos el vector $\boldsymbol{p(t)}^T = (p_1(t), p_2(t), \ldots, p_N(0))$ con elementos $p_i(t)$ la probabilidad de estar en el nodo i después de haber avanzado t pasos. Tras estos preámbulos, podemos escribir la ecuación del tiempo en el algoritmo Page Rank como la siguiente expresión:

$$p_i(t+1) = \sum_{j=1}^{N} p_j(t)\overline{\omega}_{ij} = \sum_{j=1}^{N}(1-\theta)p_j(t)\overline{\omega}_{ij} + \frac{\theta}{N}\sum_{j=1}^{N} p_j(t) = \sum_{j=1}^{N}(1-\theta)p_j(t)\overline{\omega}_{ij} + \frac{\theta}{N} \tag{3.2}$$

de tal forma que en forma matricial tenemos la misma ecuación que en la sección 2.4 y siendo $\mathbf{1_N}^T = (\underbrace{1, \ldots, 1}_{N})$,

$$\boldsymbol{p(t+1)} = (1-\theta)W\boldsymbol{p(t)} + \frac{\theta}{N}\mathbf{1_N}. \tag{3.3}$$

Para el estudio posterior también estamos interesados en cómo cambian probabilidades de los nodos, esto lo expresamos como la resta de los vectores de probabilidad entre el tiempo $t+1$ y el tiempo t, usando la ecuación 3.3 nos queda

$$\boldsymbol{p(t+1)} - \boldsymbol{p(t)} = -\left[\boldsymbol{I} - (1-\theta W)\right]\boldsymbol{p(t)} + \frac{\theta}{N}\mathbf{1_N} = -Q\boldsymbol{p(t)} + \frac{\theta}{N}\mathbf{1_N} \tag{3.4}$$

con $Q = \boldsymbol{I} - (1-\theta)W$ e $\boldsymbol{I}$ la matriz identidad.

Para terminar con esta sección, queremos ver la distribución estacionaria $\boldsymbol{\pi}$ que cumplirá

$$\boldsymbol{\pi} = (1-\theta)W\boldsymbol{\pi} + \frac{\theta}{N}\mathbf{1_N} \tag{3.5}$$

y despejando $\boldsymbol{\pi}$ nos deja la expresión

$$\boldsymbol{\pi} = \left[\boldsymbol{I} - (1-\theta)W\right]^{-1}\frac{\theta}{N}\mathbf{1_N}. \tag{3.6}$$

3.1.2. Difusión de la probabilidad

Ahora que tenemos las ecuaciones a partir de las cuales vamos a desarrollar la difusión de la probabilidad consideramos las propiedades de la matriz W. Si bien es cierto que no es simétrica, sí que es semejante a la matriz simétrica $M = G^{-\frac{1}{2}}WG^{-\frac{1}{2}}$ que gracias al los teoremas 2.1 y 2.3 podemos afirmar que tiene N autovalores $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_N > -1$ siempre y cuando $\mathcal{G}$ no sea bipartito. Debido a que W es semejante a una matriz simétrica, se puede diagonalizar y la podemos expresar como

$$W = T\Lambda T^{-1} \tag{3.7}$$

con $\Lambda = diag(\lambda_1, \cdots, \lambda_N)$. Si denotamos como $\boldsymbol{\varrho_i}$ a los autovectores izquierdos y $\boldsymbol{\varphi_i}$ a los autovectores derechos tenemos que

$$W\boldsymbol{\varphi_i} = \lambda_i \boldsymbol{\varphi_i} \text{ y } \lambda_i \boldsymbol{\varrho_i}^T = \boldsymbol{\varrho_i}^T W \tag{3.8}$$

por tanto, las matrices T y T^{-1} se definen como

$$T = (\boldsymbol{\varphi_1} | \ldots | \boldsymbol{\varphi_N}) \tag{3.9}$$

y

$$T^{-1} = \begin{pmatrix} \boldsymbol{\varrho_1}^T \\ -- \\ \vdots \\ -- \\ \boldsymbol{\varrho_N}^T \end{pmatrix} \tag{3.10}$$

Si T_{ij} es el elemento i, j de T y τ_{ij} el de T^{-1}, $T_{ij} = \boldsymbol{e_i}^T \boldsymbol{\varphi_j} = \varphi_{ij}$ y $\tau_{ij} = \boldsymbol{\varrho_i}^T \boldsymbol{e_j} = \varrho_{ij}$.Tenemos

$$\delta_{ij} = (T^{-1}T)_{ij} = \sum_k \tau_{ik} T_{k,j} = \sum_k \tau_{ik} \varphi_{jk} = \boldsymbol{\varrho_j}^T \boldsymbol{\varphi_i} \tag{3.11}$$

y

$$\delta_{ij} = (TT^{-1})_{i,j} = \sum_k T_{ik} \tau_{kj} = \sum_k \varphi_{ik} \tau_{jk} \tag{3.12}$$

La matriz W es la matriz de transición de un paseo aleatorio estándar, y sabemos que el único punto de equilibrio es el vector de probabilidad $\boldsymbol{\pi}$ con entradas $\pi_i = \frac{g(i)}{\sum_{k=1}^N g(k)}$. Por definición, $W\boldsymbol{\pi} = \boldsymbol{\pi}$, podemos deducir que el vector $\boldsymbol{\pi}$ es un autovector con autovalor asociado $\lambda = 1$ y podemos afirmar que $\boldsymbol{\varphi_1}$ es proporcional a $\boldsymbol{\pi}$ llegando a la siguiente igualdad

$$\varphi_{1i} = bg(i) \tag{3.13}$$

con $b > 0$. En un camino aleatorio estándar, en el cual si estamos en un nodo en el tiempo t , al tiempo $t + 1$ tendremos que estar en un nodo vecino del que partíamos. Si la probabilidad de moverse del nodo i al nodo j es ω_{ij} se sigue que

$$\sum_{j=1}^{N} \omega_{ij} = 1 \quad \forall i = 1, \ldots, N \tag{3.14}$$

y por eso,

$$\boldsymbol{1_N}^T W = \boldsymbol{1_N}^T \tag{3.15}$$

además como podemos sabemos que $\boldsymbol{\varrho_1}$ es el autovector izquierdo asociado al autovalor $\lambda = 1$,tiene que ser proporcional al vector de unos

$$\boldsymbol{\varrho_1} = a\mathbf{1_N} \tag{3.16}$$

con $a > 0$. A partir de las igualdades 3.14 y 3.16 llegamos a una igualdad que depende de dos parámetros a y b

$$1 = \boldsymbol{\varrho_1}^T \boldsymbol{\varphi_1} = ab \sum_{i=1}^{N} g(i) \tag{3.17}$$

Si elegimos $b = \frac{1}{\sum_{i=1}^{N} g(i)}$ nos lleva a

$$\varphi_{1,i} = \frac{g(i)}{\sum_{i=1}^{N} g(i)} \tag{3.18}$$

y $a = 1$, nos dice que $\boldsymbol{\varrho_1} = \mathbf{1_N}$. Son relevantes las siguientes propiedades

$$\sum_{k=1}^{N} \varphi_{1k} = 1 \tag{3.19}$$

$$\sum_{k=1}^{N} \varrho_{1k} = N \tag{3.20}$$

Los demás autovectores tienen autovalores asociados λ con $|\lambda| < 1$ y cuentan con una propiedad relevante.

Teorema 3.1. *Los autovectores $\boldsymbol{v}$ de la matriz W con autovalores asociados $|\lambda_i| \neq 1$ cumplen*

$$\sum_{i=1}^{N} v_i = 0 \tag{3.21}$$

Demostración. La ecuación de los autovectores nos dice

$$\sum_{i=1}^{N} \omega_{ij} v_i = \lambda v_j \tag{3.22}$$

y sumando componente a componente de los vectores nos queda

$$\lambda \sum_{j=1}^{N} v_j = \sum_{j=1}^{N} \sum_{i=1}^{N} \omega_{ij} v_i = \sum_{i=1}^{N} \underbrace{\sum_{j=1}^{N} \omega_{ij}}_{=1} v_i = \sum_{i=1}^{N} v_i \tag{3.23}$$

y como $0 \neq \lambda \neq 1$ esta igualdad solo se puede dar si

$$\sum_{i=1}^{N} v_i = 0 \tag{3.24}$$

acabando así la demostración. □

Ahora vamos a proceder a estudiar la matriz Q, por su definición en 3.4 comparte autovectores con la matriz W y tiene como autovalores

$$\mu_i = 1 - (1-\theta)\lambda_i \tag{3.25}$$

con

$$\theta = \mu_1 < \mu_2 \leq \cdots \leq \mu_N = 1 - (1-\theta)\lambda_N < 2 - \theta \tag{3.26}$$

Si consideramos el caso en el que $\theta \neq 0$, Q se puede descomponer como

$$Q = TLT^{-1} \tag{3.27}$$

con la matriz

$$L = diag(\mu_1, \ldots, \mu_N) \tag{3.28}$$

Nota: Es relevante mencionar que los autovalores en la matriz L están ordenados de menor a mayor en vez de mayor a menor como hacíamos en la matriz Λ. Esto lo hacemos para mantener la correspondencia de los autovalores λ_i y μ_i.

Antes de continuar es recomendable comprobar que nuestra distribución estacionaria es, en efecto, una distribución de probabilidad, esto es, que $\sum_{i=1}^{N} \pi_i = 1$. Por la expresión 3.6 tenemos que

$$\boldsymbol{\pi} = Q^{-1}\frac{\theta}{N}\mathbf{1_N} = TL^{-1}T^{-1}\frac{\theta}{N}\mathbf{1_N} \tag{3.29}$$

Por la dimensión de T^{-1},

$$(T^{-1}\mathbf{1_N})_j = \sum_{k=1}^{N} \tau_{jk} \tag{3.30}$$

por tanto,

$$T^{-1}\mathbf{1_N} = \sum_{h=1}^{N}\sum_{k=1}^{N} \tau_{hk}\boldsymbol{e_h} \tag{3.31}$$

de aquí se sigue

$$L^{-1}T^{-1}\mathbf{1_N} = \sum_{h=1}^{N}\frac{1}{\mu_h}\sum_{k=1}^{N} \tau_{hk}\boldsymbol{e_h} \tag{3.32}$$

y

$$TL^{-1}T^{-1}\mathbf{1_N} = \sum_{h=1}^{N} \frac{1}{\mu_h} \sum_{k=1}^{N} \tau_{hk}\boldsymbol{\varphi_h} \tag{3.33}$$

en consecuencia,

$$\boldsymbol{\pi} = \frac{\theta}{N} \sum_{h=1}^{N} \frac{1}{\mu_h} \sum_{k=1}^{N} \tau_{hk}\boldsymbol{\varphi_h} \tag{3.34}$$

$$\sum_{i=1}^{N} \pi_i = \frac{\theta}{N} \sum_{h=1}^{N} \frac{1}{\mu_h} \sum_{k=1}^{N} \tau_{hk} \sum_{i=1}^{N} \varphi_{hi}. \tag{3.35}$$

Por el teorema 3.1 todos los sumandos se anulan salvo para $h = 1$ quedando

$$\sum_{i=1}^{N} \pi_i = \frac{\theta}{N} \frac{1}{\mu_1} \sum_{k=1}^{N} \tau_{1k}\varphi_{1i} = \frac{\theta}{N} \frac{1}{\theta} N = 1 \tag{3.36}$$

ahora que ya hemos comprobado que $\boldsymbol{\pi}$ es efectivamente un vector de probabilidad, volvemos a la ecuación 3.4.

Si definimos el incremento de tiempo Δt y asumimos que el lado derecho de la ecuación es constante, obtenemos

$$\boldsymbol{p(t + \Delta t)} - \boldsymbol{p(t)} = \Delta t[-Q\boldsymbol{p(t)} + \frac{\theta}{N}\mathbf{1_N}]. \tag{3.37}$$

Para considerar t como una variable continua dividimos a ambos lados por Δt y hacemos el límite $\Delta t \to 0^+$ obteniendo así la aproximación continua del camino, esto es,

$$\frac{\mathrm{d}\boldsymbol{p(t)}}{\mathrm{d}t} = -Q\boldsymbol{p(t)} + \frac{\theta}{N}\mathbf{1_N}. \tag{3.38}$$

Esta ecuación puede ser interpretada como una ecuación de difusión bajo el operador Q. Estamos interesados en estudiar la evolución en el tiempo del vector de probabilidades según nos aproximamos a la distribución estacionaria. La solución al sistema de ecuaciones ordinarias queda como

$$\boldsymbol{p(t)} = e^{-Qt}C + \frac{\theta}{N}Q^{-1}\mathbf{1_N}. \tag{3.39}$$

donde C es un vector que depende de la distribución de probabilidad inicial $\boldsymbol{p(0)}$ si lo resolvemos queda,

$$\boldsymbol{p(t)} = e^{-Qt}\boldsymbol{p(0)} + \frac{\theta}{N}(\boldsymbol{I} - e^{-Lt})Q^{-1}\mathbf{1_N} = Te^{-Lt}T^{-1}\boldsymbol{p(0)} + \frac{\theta}{N}T(\boldsymbol{I} - e^{-Lt})L^{-1}T^{-1}\mathbf{1_N} \tag{3.40}$$

3.1.3. Estudio del número medio de vecinos en un grafo de tipo leskovec

En la sección 3.1.4 aparecerá que el número de nodos visitados en el algoritmo de Page Rank depende del número de vecinos medio del grafo. Concretamente, en el tipo de grafos que hemos utilizado el número de vecinos crece como una distribución potencia de parámetro γ, esto nos dice que

$$P(X < x) = \sum_{x=1}^{\infty} c\frac{1}{x^{\gamma}} \tag{3.41}$$

Por tanto, el número medio de vecinos $\bar{q}$ será la esperanza de esta distribución de probabilidad.

$$E(X) = \sum_{x=1}^{\infty} xP(X = x) \tag{3.42}$$

Volviendo a nuestro estudio concreto, será

$$\bar{q} = \sum_{x=1}^{\infty} c\frac{1}{x^{\gamma-1}} \tag{3.43}$$

Esta es una función zeta de Riemman que dificulta el estudio, por esto, añadiremos la restricción de que el número máximo de vecinos en el grafo debe ser finito, esta restricción es natural ya que los grafos sobre los que estamos trabajando tienen un número finito de nodos ($\bar{k}$). Nos queda por tanto,

$$\bar{q} = \sum_{x=1}^{\bar{k}} c\frac{1}{x^{\gamma-1}} \tag{3.44}$$

que es fácilmente computable.

3.1.4. Estudio del número de nodos visitados en función del tiempo en el Page Rank

El problema que nos atañe se centra en estudiar cuán de rápido podemos visitar la totalidad de los nodos del grafo .
Consideremos que el grafo cuyo número medio de vecinos por nodo lo denotamos como $\bar{q}$, por tanto, el caminante aleatorio al viajar a través de una arista de media nos lleva a un nodo de $\bar{q}$ vecinos y, si $m(t)$ es es número de nodos distintos visitados al tiempo t, la probabilidad de visitar un nodo el cual hayamos visitado será $\frac{m(t)}{N}$ y la de visitar un nodo por el que no hayamos pasado será $1 - \frac{m(t)}{N}$.

Supongamos ahora que acabamos de alcanzar el nodo n en el grafo. Existen dos formas de haberlo alcanzado: o bien venimos de un nodo $\bar{n}$ que es vecino de n (esto ocurre con probabilidad $1 - \theta$), o bien hemos decidido saltar desde un nodo aleatorio y hemos caído en n (esto ocurre con probabilidad θ). Una vez que estamos en n volvemos a tener dos alternativas, desplazarnos a un vecino con pro-

babilidad $1-\theta$ o saltar a un nodo aleatorio del grafo con probabilidad θ. Vamos ahora a considerar las dos formas de alcanzar en nodo n y las dos maneras de abandonarlo.

- Si hemos llegado desde un vecino: el nodo n tiene de media $\bar{q}$ vecinos y entre estos está $\bar{n}$. Para abandonarlo volvemos a lanzar una θ-moneda que da lugar a dos casuísticas
  - Si nos movemos a un vecino del nodo n: la media de vecinos es $\bar{q}$, por tanto lo probabilidad de volver al nodo desde el que llegamos a n es $\frac{1}{\bar{q}}$. Si nos movemos a cualquier otro vecino (con probabilidad $1-\frac{1}{\bar{q}}$ la probabilidad de que no haya sido visitado es de $1-\frac{m(t)}{N}$. Así llegamos a la conclusión de que si el caminante sigue esta vía, la probabilidad de de visitar un nuevo nodo queda

$$(1-\theta)^2(1-\frac{1}{\bar{q}})(1-\frac{m(t)}{N}) \tag{3.45}$$

  - Si vamos a un nodo aleatorio del grafo: en este caso no tiene importancia que uno de nuestros vecino ya haya sido visitado, saltamos a un nodo del grafo con probabilidad de que no haya sido visitado $1-\frac{m(t)}{N}$. Si hacemos la conjunción de todos las probablidades que han tenido que darse para este caso, tenemos que la probabilidad de saltar a un nodo por el cual no hayamos pasado es

$$(1-\theta)\theta(1-\frac{m(t)}{N}) \tag{3.46}$$

- Si hemos llegado a n desde un nodo aleatorio del grafo no importa la forma en la cual abandonemos el nodo. Estamos en una zona del grafo de la cual no tenemos información, nuestros vecinos son nodos cualquiera. Por tanto tenemos que la probabilidad de visitar un nodo nuevo es

$$\theta(1-\frac{m(t)}{N}) \tag{3.47}$$

Si ponemos todas las probabilidades juntas tenemos que en un paso de tiempo la probabilidad de que el nodo al que nos desplazamos no haya sido visitado es:

$$(1-\theta)\left[(1-\theta)(1-\frac{1}{\bar{q}})(1-\frac{m(t)}{N})+\theta(1-\frac{m(t)}{N})\right]=\theta(1-\frac{m(t)}{N})= \tag{3.48}$$

$$=\left[(1-\theta)^2(1-\frac{1}{\bar{q}})+(1-\theta)\theta+\theta\right](1-\frac{m(t)}{N})=\nu(1-\frac{m(t)}{N}) \tag{3.49}$$

donde ν depende de θ y puede ser escrito como

$$\nu=-\frac{1}{\bar{q}}\theta^2+\frac{2}{\bar{q}}+(\frac{\bar{q}-1}{\bar{q}}) \tag{3.50}$$

Si escribimos el incremento medio del número de nodos visitados en el camino aleatorio del Page Rank queda

$$m(t+1)-m(t)=\nu(1-\frac{m(t)}{N}) \tag{3.51}$$

y exigiendo un incremento temporal entre pasos infinitesimal obtenemos la derivada de $m(t)$

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t}=\nu(1-\frac{m(t)}{N}) \tag{3.52}$$

que al resolverla imponiendo la condición inicial $m(0) = 1$

$$m(t) = N(1 - e^{-\frac{\nu t}{N}}) + e^{-\frac{\nu t}{N}} \tag{3.53}$$

en la imagen 3.1 se ve cómo crece el número de nodos visitados en función del tiempo según vamos cambiando el parámetro $\beta = 1 - \theta$

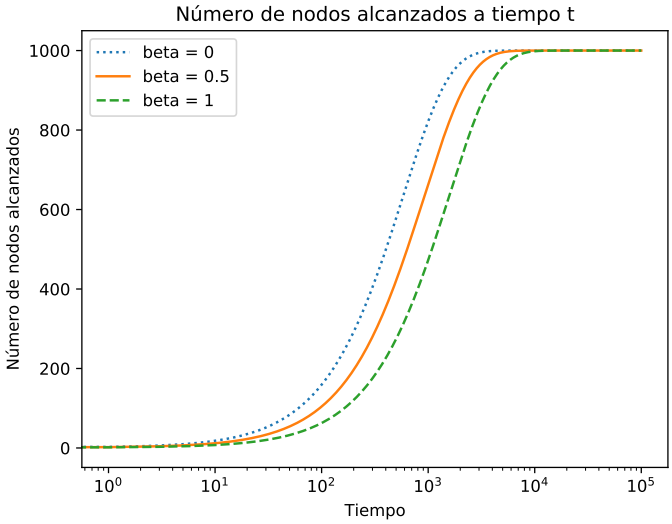

Figura 3.1: Crecimiento teórico del número de nodos visitados en función del tiempo en el Page Rank

Para terminar con esta sección es importante añadir un par de apuntes. Esta dinámica es muy similar a la probabilidad de difusión y ambas están relacionadas. La difusión es más rápida cuando ν es grande que crece en forma de polinomio cuadrático en fución de β y alcanza su máximo cuando $\beta = 1$ (difusión totalmente aleatoria). Estos resultados los veremos de manera empírica en la sección 3.2. Sin embargo, en casos prácticos, esto no significa que la búsqueda totalmente aleatoria de información sea la mejor estrategia de búsqueda. Se debe tener en cuenta que los sitios cercanos suelen tener información relacionada por lo que una estrategia de búsqueda que combine saltos totalmente aleatorios del Page Rank y una búsqueda en un área restricta sería un buen candidato a a ser una estrategia óptima. Esto escapa del alcance de este trabajo y se deja como problema abierto a explorar.

3.1.5. Tiempo medio entre nodos

Ahora que tenemos la velocidad a la que el paseo aleatorio explora la totalidad del grafo queremos ver cuánto tarda de media en llegar del nodo j habiendo empezado en el nodo i. Un punto relevante a mencionar es que las probabilidades discretas de estancia en los nodos al tiempo t, escritas como $p_{ii}(t)$ se transforman en densidades de probabilidad cuando el tiempo entra en juego, gracias a este hecho si empezamos el camino aleatorio en el nodo i, la probabilidad de llegar al nodo j en el intervalo

de tiempo $[t, t+\Delta t]$ es $p_{ij}(t)\Delta t$ permitiéndonos escribir

$$p_{ij} = \delta_{ij}\delta(0) + \int_0^t p_{jj}(t-\tau)F_{ij}(\tau)\mathrm{d}\tau \tag{3.54}$$

Donde δ_{ij} es la delta de Kronecker, $\delta(0)$ es la de Dirac y $F_{ij}(t)$ es la probabilidad de primera pasada, esto es, $F_{ij}(t)\Delta t$ es la probabilidad de, empezando por el nodo i hemos visitado por primera vez el nodo j en el intervalo de tiempo $[t, t+\Delta t]$, si a la ecuación del tiempo 3.54 la aplicamos la transformada de Laplace continua definida como $\tilde{g}(s) = \int_0^\infty e^{-st}\mathrm{d}t$ queda

$$\tilde{p}_{ij}(s) = \delta_{ij} + \tilde{p}_{jj}(s)\tilde{F}_{ij}(s) \tag{3.55}$$

tenemos así una ecuación que nos despeja $\tilde{F}_{ij}(s)$

$$\tilde{F_{ij}}(s) = \frac{\tilde{p}_{ij}(s) - \delta_{ij}}{\tilde{p}_{jj}(s)} \tag{3.56}$$

que está relacionada con el tiempo esperado de ir del nodo i al nodo j ya que este se expresa como la media de las probabilidades de primera pasada

$$\langle T_{ij}\rangle = \int_0^\infty tF_{ij}(t)\mathrm{d}t = -\frac{\mathrm{d}\tilde{F}_{ij}(0)}{\mathrm{d}s} \tag{3.57}$$

Para obtener la expresión del lado derecho de la ecuación 3.57 vamos a extender $\tilde{p_{ij}}(s)$

$$\tilde{p}_{ij}(s) = \int e^{-st}p_{ij}(t)dt = \pi_j\int e^{-st} + \int e^{-st}\big[p_{ij}(t) - \pi_j\big]dt = \pi_j\int e^{-st} + \int e^{-st}\hat{p}_{ij}(t)dt \tag{3.58}$$

con $\hat{p}_{ij}(t) = p_{ij}(t) - \pi_j$. Calculando la primera integral y expresando la exponencial como una serie obtenemos

$$\tilde{p}_{ij}(s) = \frac{\pi_j}{s} + \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}s^n\int t^n\hat{p}_{ij}(t)dt = \frac{\pi_j}{s} + \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}s^nR_{ij}^n = \frac{\pi_j}{s} + Q_{ij}(s) \tag{3.59}$$

donde definimos los momentos

$$R_{ij}^n = \int_0^\infty t^n\hat{p}_{ij}(t)dt \tag{3.60}$$

y la abreviatura Q_{ij}

$$Q_{ij}(s) = \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}s^nR_{ij}^n \tag{3.61}$$

Vemos que $Q_{ij}(0) = R_{ij}^{(0)}$. Gracias a ello, podemos escribir 3.56 como

$$\tilde{F}_{ij}(s) = \frac{1}{\frac{\pi_j}{s} + Q_{jj}(s)}\Big[\frac{\pi_j}{s} + Q_{ij}(s) - \delta_{ij}\Big] = \frac{\pi_j + s(Q_{ij}(s) - \delta_{ij})}{\pi_j + sQ_{jj}(s)} \tag{3.62}$$

y tiene como derivada,

$$\tilde{F}'_{ij}(s) = \frac{Q_{ij}(s) - \delta_{ij} + sQ^T_{ij}(s)}{\pi_j + sQ_{jj}(s)} - \frac{(Q_{jj}(s) + sQ^T_{jj}(s))\big[\pi_j + s(Q_{ij}(s) - \delta_{ij})\big]}{(\pi_j + sQ_{jj}(s))^2} \tag{3.63}$$

que al evaluarla en $s = 0$ queda

$$\langle T_{ij}\rangle = \frac{1}{\pi_j}\big[R^{(0)}_{jj} - R^{(0)}_{ij} + \delta_{ij}\big] \tag{3.64}$$

si la evaluamos ahora para todo par de nodos

$$\langle T\rangle = \sum_{i\neq j}\langle T_{ij}\rangle\pi_j = \sum_{i=1}^{N} R^{(0)}_{ii} \tag{3.65}$$

para calcular los valores de $R^{(0)}_{ij}$ tenemos que expresar $p_{ij}(t)$. y de la solución del sistema de EDOS tenemos

$$p_{ij}(t) = \frac{\theta}{N}\boldsymbol{e}_j^T(\boldsymbol{I} - e^{-Qt})Q^{-1}\mathbf{1}_N + \boldsymbol{e}_j^T e^{-Qt}\boldsymbol{e}_i \tag{3.66}$$

lo que implica

$$\hat{p}_{ij}(t) = \boldsymbol{e}_j^T e^{-Qt}\boldsymbol{e}_i - \frac{\theta}{N}\boldsymbol{e}_j^T e^{-Qt}Q^{-1}\mathbf{1}_N = \boldsymbol{e}_j^T T e^{-Lt}T^{-1}\boldsymbol{e}_i - \frac{\theta}{N}\boldsymbol{e}_j^T T e^{-Lt}L^{-1}T^{-1}\mathbf{1}_N \tag{3.67}$$

Usando las definiciones de T, T^{-1} y definiendo, $\boldsymbol{b^T} = (\sum_{k=1}^N \tau_{1k}, \sum_{k=1}^N \tau_{2,k}, \ldots, \sum_{k=1}^N \tau_{N,k})$ podemos desarrollar la expresión como

$$\hat{p}_{ij}(t) = \sum_{k=1}^{N} e^{-\mu_k t}\varphi_{kj}\tau_{ki} - \frac{\theta}{N}\frac{e^{-\mu_k t}}{\mu_k}\varphi_{kj}b_k \tag{3.68}$$

gracias a 3.68 podemos escribir los momentos como

$$R^{(0)}_{ij} = \int_0^\infty \hat{p}_{ij}(t) = \sum_{k=1}^{N}\frac{1}{\mu_k}\Big[\varphi_{kj}\tau_{ki} - \frac{\theta}{N\mu_k}\varphi_{kj}b_k\Big] = \sum_{k=1}^{N}\frac{1}{\mu_k}\tau_{ki}\tau_{ki} - \frac{\theta}{N\mu_k^2}\sum_{k=1}^{N}\varphi_{ki}b_k \tag{3.69}$$

y aplicando la igualdad 3.65 tenemos una forma de escribir el tiempo medio de llegada entre nodos

$$\langle T\rangle = \sum_{i=1}^{N}\sum_{k=1}^{N}\frac{1}{\mu_k}\varphi_{ki}\tau_{ki} - \sum_{i=1}^{N}\sum_{k=1}^{N}\frac{\theta}{N\mu_k^2}\varphi_{ki}b_k = \sum_{k=1}^{N}\frac{1}{\mu_k}\sum_{i=1}^{N}\varphi_{ki}\tau_{ki} - \sum_{k=1}^{N}\frac{\theta}{N\mu_k^2}b_k\sum_{i=1}^{N}\varphi_{ki} \tag{3.70}$$

El primer sumatorio que depende de i es 1 gracias a la propiedad 3.12, en el segundo sumatorio los últimos términos son igual a 0 por el teorema 3.1. Para terminar, si tenemos en cuanta que $\mu_1 = \theta$ llegamos a la expresión

$$\langle T\rangle = \sum_{k=1}^{N}\frac{1}{\mu_k} - \frac{u}{N\mu_1^2}\sum_{i=1}^{N}\varphi_{1i}\sum_{i=1}^{N}N\tau_{k1} = \sum_{k=1}^{N}\frac{1}{\mu_k} - \frac{1}{\mu_1} = \sum_{k=2}^{N}\frac{1}{\mu_k} = \sum_{k=2}^{N}\frac{1}{1-(1-\theta)\lambda_k} \tag{3.71}$$

Si comparamos esta expresión con la provista en [4] y tomamos $\theta = 0$ vemos que son la misma que en un camimo de Levy de parámetro $\alpha \to \infty$. Un hecho relevante a mencionar es que en los autovalores

de [4] la dependencia del parámetro α está oculta en los autovalores λ_k mientras que en nuestro caso, el Page Rank, estos autovalores solo dependel de la estructura del grafo y la dependencia de θ es explícita.

3.2. Comparación del número de nodos visitados de Page Rank y un camino de Lévy

Para el estudio de la difusión creamos un grafo a partir del algoritmo explicado en la sección 2.1 bajo los parámetros:

- **Número de nodos:** 1000
- **Tiempo de vida de los nodos(δ):** 1
- **Probabilidad de cerrar un triángulo:** 0.2

Estos parámetros se escogieron de esta forma debido a que bajo los valores sugeridos por [5] incluso con mil nodos el diámetro era demasiado pequeño y no se apreciaban los cambios de parámetro del paseo de Lévy y el Page Rank. Una vez hecho esto, se estudiaron varias propiedades de los paseos aleatorios:

- probabilidad de estar a una distancia en función del tiempo
- tiempo necesario para alcanzar una distancia d

Estas simulaciones se realizaron con la idea de relacionar el Page Rank con los caminos de Lévy, sin embargo, los resultados no eran concluyentes, la aleatoriedad del paseo afectaba a la distancia a la que estábamos, ya que si escogíamos el nodo de partida como referencia, la distribución que hallábamos según aumentaba el tiempo variaba de forma arbitraria, llegamos a la conclusión de que aunque existiera una distribución estacionaria que dependía de cada nodo, podría darse el caso de que dos nodos a la misma distancia del nodo de referencia tuvieran probabilidades muy dispares lo que añadía ruido a nuestros resultados.

Esto nos hizo tomar la perspectiva de [4] de observar el número de nodos cubiertos por el paseo de Lévy o algoritmo Page Rank. En las figuras 3.2 y 3.3 se muestra la gráfica de varias simulaciones realizadas, sus características son: tomando un grafo de 1000 nodos queríamos ver el tiempo necesario para que el caminante aleatorio pasara por todos los nodos, esto se hizo 50 veces para cada valor de α y β llevando la cuenta del número de nodos que el caminante había visitado, los puntos en las gráficas son la media para cada tiempo, es relevante que en ambas apreciamos mucha similitud. Las gráficas correspondientes a $\alpha = 0$ y $\beta = 0$ son las gráficas que cubren todo el grafo en el menor número de pasos en el tiempo lo que concuerda con la figura 3.1 que mostraba nuestras expectativas teóricas para el PageRank, esto ocurre porque cómo se analizó en la sección 2 el paseo aleatorio puede saltar a cualquier nodo en cada paso de tiempo y aunque teóricamente, son una buena estrategia de bús-

queda de información para llegar al óptimo deberían de combinarse con algoritmos de búsqueda local, además vemos cómo se confirmaban nuestras sospechas de que la "perezosidad" de este caso límite en el Page Rank no es significativa por ser en este caso de $\frac{1}{1000}$. Si observamos el caso intermedio de un camino de Lévy, $\alpha = 3$, vemos que el tiempo necesario para cubrir todo el grafo es algo superior al caso anteriormente mencionado, este fenómeno es debido a la probabilidad decreciente de salto a un nodo según aumenta la distancia. En el caso del Page Rank $\beta = 0{,}5$ estamos moviéndonos a nodos vecinos una de cada dos veces que nos desplazamos, el restos de veces, estamos visitando nodos vecinos como si en un paseo alteatorio estándar de tratara. Por último tenemos el recorrido del grafo de forma tradicional $\alpha = 100$ y $\beta = 1$, en el cual el caminante aleatorio solo puede desplazarse a nodos que estén a distancia uno, esta limitación tiene un gran impacto a la hora de tratar de visitar todos los nodos llegando a alcanzar valores del orden de 10^5 mientras que en el caso más eficiente la totalidad del grafo se cubre en un tiempo de alrededor de 10^4 incrementos.

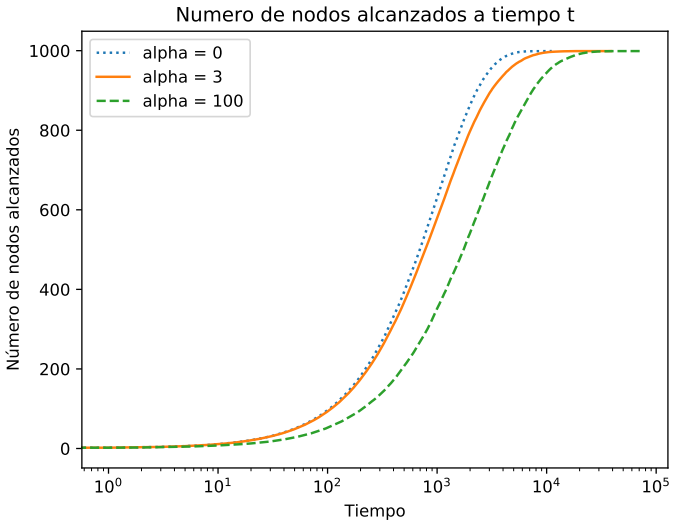

Figura 3.2: Simulación de nodos visitados por tiempo en un camino de Lévy

Una vez hecho este análisis procedemos a estudiar las expresiones obtenidas en la sección 3.1.4 y las compararmos con datos empíricos, para este tipo de simulaciones escogíamos dos nodo cualquiera del grafo asignábamos a uno de ellos como nodo inicial y al otro como nodo destino, tras esto, realizábamos una simulación del paseo aleatorio correspondiente que finalizábamos al llegar al nodo destino, almacenábamos el tiempo tardado y volvíamos a proceder de la misma forma 200 veces, para obtener el tiempo medio, calculamos la media de todos estos tiempos, estas son las gráficas punteadas de color azul que aparecen en las figuras 3.3 y en 3.2, para los tiempos teóricos nos valíamos de las fórmulas 3.71 y 2.38 respectivamente para calcularlos.

Como vemos en la gráfica correspondiente al camino de Lévy, 3.4 el tiempo medio entre nodos tiene aumenta cuando incrementamos el parámetro α. La disparidad de los datos teóricos frente a los empíricos se debe a que, cuando procedemos a realizar un camino aleatorio en particular, pueden

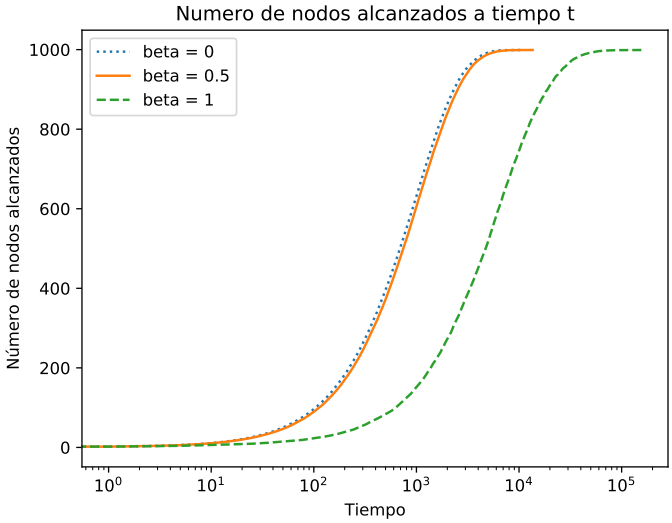

Figura 3.3: Simulación de nodos visitados por tiempo en en el Page Rank

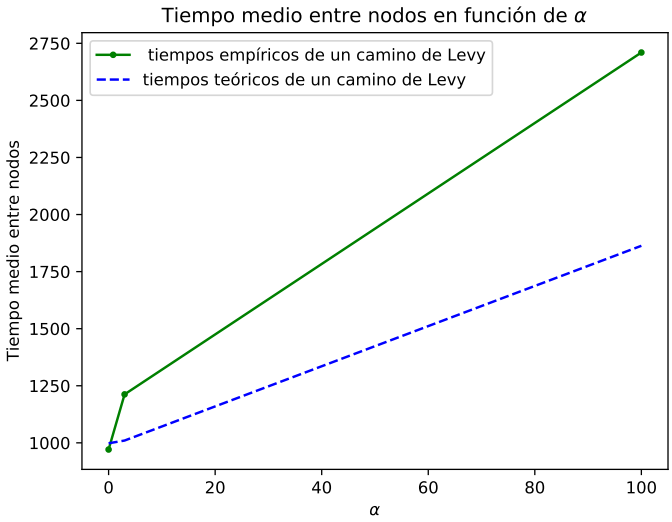

Figura 3.4: Simulación del tiempo medio entre nodos en caminos de Lévy

existir datos atípicos que nos desvíen la media de las simulaciones respecto a el cálculo teórico, sin embargo la tendencia de las gráficas confirma que los resultados provistos por [4] sean fiables.

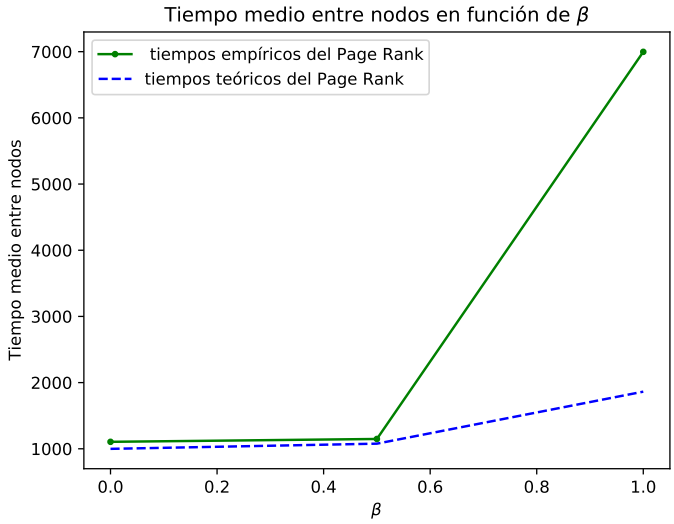

Figura 3.5: Simulación del tiempo medio entre nodos en el Page Rank

El último conjunto de simulaciones realizado se centró en el tiempo medio entre nodos del algoritmo Page Rank tomando como punto de partida la ecuación expuesta en 3.71 en ella vemos que según aumenta el parámetro β, los tempos teóricos se corresponden con los empíricos en los casos $\beta = 0$ y $\beta = 0{,}5$ con una gran disparidad en el caso $\beta = 1$, esta disparidad la atribuimos a las simulaciones que pueden tener una gran desviación típica.

4

Conclusión

Para terminar recordemos que estamos estudiando la difusión en forma de paseos aleatorios, un ejemplo de la vida cotidiana sería un usuario saltando de perfil en perfil buscando información de, por ejemplo, un cotilleo. Otra forma de verlo sería que nosotros estamos estáticos y nos llegan noticias por difusión de usuario a usuario, en este segundo caso será la información el elemento que modeliza el paseo aleatorio.

Si recopilamos todos los resultados obtenido en las anteriores secciones, vemos un comportamiento similar en el Page Rank y en los paseos de Lévy, sus gráficas de nodos visitados respecto al tiempo tienen mucha similitud, los resultados teóricos que hemos obtenido a lo largo del capítulo 3 del Page Rank con parámetro $\beta = 1$ nos dan una correspondencia con [4] con el parámetro $\alpha \to \infty$. Si nos fijamos en el tiempo entre nodos vemos que en las gráficas que tienen una tendencia ascendente con un orden similar de tiempos en α y β iguales a 0 . Los tiempos medios entre nodos y las gráficas de recubrimiento de la totalidad del grafo en función del tiempo, nos dan pruebas empíricas. Sin embargo, si nos fijamos en la construcción de sus ecuaciones, el Page Rank es más fiel a la realidad. Recurriendo al ejemplo de un usuario navegando por la red que está buscando información sobre un tema en una red social, el comportamiento natural es entrar en un perfil y empezar nuestro paseo aleatorio clickando en enlaces que nos llevan de un perfil a otro, esta parte sería el equivalente a la parte de un paseo aleatorio estándar, en esta situación sólo podemos movernos a nodos de distancia uno ya que necesariamente necesitamos pasar por una página puente para llegar a un link que actualmente no vemos, existe una probabilidad $1 - \beta$ de que este usuario se canse de este sistema ya que, por lo general, los perfiles vecinos contienen mayoritariamente la misma información, entonces, este usuario utilizará el buscador y le dará un conjunto de perfiles que no tienen por qué estar referenciados desde la página que partimos, este hecho lo modela el Page Rank como la probabilidad $1 - \beta$ de saltar a los nodos de forma equiprobable. Por otro lado, en un paseo de Lévy, la probabilidad de saltar a un nodo del grafo decrece según aumenta la distancia, cosa que no ocurre cuando un individuo está navegando por la red ya que los motores de búsqueda, las páginas que nos ofrecen no dependen de la distancia a la que nos encontremos de ellas. Ésta es la principal diferencia entre ambos algoritmos y el porqué el Page Rank es una modelización más fiel de lo que ocurre en la red.

Bibliografía

[1] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, pp. 845, Aug 2000.

[2] M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference*. Little, Brown and Company, 2000.

[3] M. Franceschet, "Pagerank: Standing on the shoulders of giants," 2010.

[4] A. P. Riascos and J. L. Mateos, "Long-range navigation on complex networks using lévy random walks," *Phys. Rev. E*, vol. 86, p. 056110, Nov 2012.

[5] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, (New York, NY, USA), p. 462–470, Association for Computing Machinery, 2008.

[6] D. A. Spielman, "Daniel a. spielman lecture notes," September 2010.


UAM
UNIVERSIDAD AUTONOMA
DE MADRID